Future Trends for the Data Fabric

*Data Fabric IG, P11*

**Over the past plenaries, the RDA Data Fabric IG has worked on issues focusing mostly on the role and exploitation of Persistent Identifiers (PIDs), originating in a discussion on essential components that are common across a wide variety of research infrastructures and research data management practices. There has been significant progress in the area of PIDs, resulting among other things in a supporting output document, the GEDE group and the Kernel Information group.**

**It is now time to revisit the original concerns of data fabric components and look at other aspects pertinent to the Data Fabric idea. The Plenary session in Berlin will be used to discuss these points and set the pathway for the future activities of the group. While we need a set of examples to feed the discussion, the list is in principle open-ended, and all contributions before and during the plenary are most welcome.**

## A common object and collection management approach (separate joint session)

*Tobias Weigel, Larry Lannom, …*

RDA has produced several outputs around digital objects, including in particular the PID Information Types and Research Data Collections specification. The PID Kernel Information WG is preparing another output that complements the Information Types recommendation. In addition, many communities have established and mature metadata schemas and services around them, surch as searching and curating.

The RDA recommendations describe common APIs for working with properties of individual objects and aggregations of multiple objects. However, the management of objects in research infrastructures can benefit from a more comprehensive management of objects, no matter whether they are singular or in collections: All of them require agreements on the most common operations, such as create, copy, move and delete, but also second-level operations with small predefined workflows, such as replication, versioning or other derivation actions.

**This is a super topic across the other contributions listed further below. It will have a dedicated joint session at the P11.**

## Object management and provenance in data analytics

*Sandro Fiore, Sofiane Bendoukha, …*

For the European Open Science Cloud (EOSC), the ENES community provides a comprehensive data analytics service called ECAS (ENES Climate Analytics Service). Originally designed for usage scenarios within the Earth system modelling community, the service is now available for further cross-disciplinary use. It is based on the concepts of datacubes, enabling efficient computation on structured, multidimensional data.

As part of future developments, ECAS can become connected to PID management services on all sides (input, output, processing, …). This opens up a chance to build a small-scale demonstrator that exploits the provenance capabilities laid out via Kernel Information. One of the FAIR aspects is re-usability, and an approach based on this exemplary analytics service, embedded within EOSC, can provide a showcase both within individual communities and across them with real users involved in the evaluation and design.

## ENVRI provenance concerns and their mapping to Data Fabric components

*Abraham Nieva de la Hidalga, Alex Hardisty, Barbara Magagna, Stephan Kindermann.*

The ENVRIplus project is analysing the data life cycle of research infrastructures to determine the points where provenance information is exposed and the types of systems components which can suport and use provenance. The data lifecycle of IS-ENES provides an example of the type of data lifecycle to be modelled. IS-ENES has identified seven steps in their data lifecycle: (1) (Raw) data generation, (2) data postprocessing/homogenisation, (3) data ingest into data centers, (4) publication into data federation, (5) versioning/errata information, (6) processing and derived data products, and (7) Long term archival and DOI assignment. In this lifecycle, generation and management of provenance data is required to help tracking the states of data products states in the different the steps of the lifecycle. Internally, the data lifecycle involves different entities that contribute to the research data lifecycle. There are six types of entities participating in IS-ENES: (1) modelling centres, (2) observation data providers, (3) data nodes, (4) index nodes, (5) compute (processing/modelling) nodes and (6) [access] portals. Modelling centres and observation data providers are primary sources for the data nodes within the ENES data infrastructure. Index nodes provide cataloguing sercives to facilitate access to data nodes. Compute nodes facilitate generation of derived data products on demand. Access portals facilitate access to data produts provided by the ENES data infrastructure. The IS-ENES lifecycle can be aligned at two levels with the proposal of the DFIG: mapping of IS-ENES components to GDOC components and support required form additional GDOC components. IS-ENES components map to two main types of GDOC objects: Repository Registries (observation data providers, data nodes) ecosystem of tools [tool registries] (modelling centres, compute nodes), and metadata registries (index nodes). In addition to aligning IS-ENES components to the GDOC, additional components are required to support the processes: PID registries to provide identifiers for the differnt data products, autorisation and license registries to manage access to collections (supporting the work of the ENES portal), and type registries to facilitate access and reuse of data by third parties. It is expected that the correspondence and support of IS-ENES components to GDOC virtual layer is a common occurence in research infrastructures within the ENVRI Community.

## Metadata components and metadata fabrics

*Rainer Stotzka, Kirsten Elger, …*

In RDA many disciplinary groups aim to develop or seek help in metadata descriptions within their scientific domains. The term “Metadata Fabric” aims to develop a methodological and technical approach to support a resourceful and technically sustainable research data management by metadata. It will invest in a systematic analysis of current metadata methodologies, technologies and standards within and outside of RDA to derive commonalities. As a long term goal a set of interoperable technical components will be recommended as a basis for a metadata infrastructure harmonizing with the components of the data fabric.

## Chinese Academy of Sciences activities

*LI Jianhui, …*

# Draft agenda

1. Introduction (10 min.)
2. Future trends – individual topics under common object and collection management (40 min.):
   1. Object management and provenance in data analytics
   2. ENVRI provenance concerns and their mapping to Data Fabric components
   3. Metadata components and metadata fabrics
   4. Activities by the Chinese Academy of Sciences
3. Open discussion (30 min.)
4. New co-chairs, next steps. (10 min.)

Slot 2 should not cover more than half of the session time. If we have more contributions, we will merge them into groups.